Neurocomputing 195 (2016) 104–111

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

$L_{2,p}$−norm and sample constraint based feature selection and classification for AD diagnosis

Mingxing Zhang [a], Yang Yang [a,*], Hanwang Zhang [b], Fumin Shen [a], Dongxiang Zhang [b]

[a] University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731 Chengdu, China
[b] National University of Singapore, 21 Lower Kent Ridge Road, 119077, Singapore

ARTICLE INFO

Article history:
Received 21 March 2015
Received in revised form
26 June 2015
Accepted 23 August 2015
Available online 1 April 2016

Keywords:
Alzheimer's Disease (AD)
Mild Cognitive Impairment (MCI)
$L_{2,1}$−norm
Sparse learning
Multi-modality learning
Multi-task learning

ABSTRACT

Recent studies have witnessed the effectiveness of $L_{2,1}$-norm based methods on AD/MCI diagnosis. Nonetheless, most of them suffer from the following three main problems: (1) $L_{2,1}$−norm based loss function does not take into account different distances between target labels and prediction values; (2) $L_{2,1}$−norm based feature selection does not possess sufficient flexibility to adapt to different types of data sources and select more informative features; (3) intrinsic correlation between the processes of feature selection and classification (or regression) are inevitably ignored. In this paper, we propose a novel method which incorporates additional flexibility and adaptability by employing the more generalized $L_{2,p}$−norm based prediction loss function and $L_{2,q}$−norm based feature selection, as well as utilizes a joint model to perform feature selection and classification simultaneously. Besides, we introduce a regularizer to preserve local structure information between samples in the original feature space and prediction values in the projected space. In order to validate the effectiveness of the proposed method, we conducted extensive experiments on the ADNI dataset, and showed that the proposed method enhanced the performance of disease status classification, compared to the state-of-the-art methods.

1. Introduction

Alzheimer's disease (AD) is the most common dementia in elderly people, which results in serious intellectual problems of memory, thinking and behavior. According to the report of Brookmeyer et al. [1], there would be over 30 million people around the world suffering from this disease by 2050. Its prodromal stage, which is called Mild Cognitive Impairment (MCI), can also lead to cognitive changes and high risk of development of AD over times [2]. It is very important to diagnose the AD/MCI and many researches about automatic computer-aid diagnosis of these diseases have been conducted in recent decades.

One of the main problems in the field of automatic medical diagnosis is that the dimension of medical data is normally far larger than the sample size. For example, the size of samples in many researches such as [3–5] was very small (only 103 samples with 51 for AD and 52 for NC), while the dimension of data features such as MRI and PET reached hundreds to thousands. However, AD is only related to a few areas of brain according to the research in [5]. These high-dimensional features usually contain many uninformative features. The high dimension of data also could result in the over-fitting problem [6] and the small size of samples makes it more serious.

To address these issues, feature selection based methods have been commonly used in literatures. As the successful applications of sparse method such as [7–11], $L_{2,1}$−norm based methods have been widely used in feature selection process for AD diagnosis. Wang et al. [12] proposed a multi-task learning method that performed the label classification and cognitive measure scores regression simultaneously. Different from traditional methods that selected features only associated with cognitive measure scores or disease status, this method selected the features related to both of them. Zhang et al. [3] proposed a multi-modal multi-task learning method that firstly selected the subset of features using Multi-Task method from each modality, then used the multi-modal support vector for the classification of AD and MCI. However, these methods do not consider the relationship between target vectors of samples. To overcome this disadvantage, Liu et al. [4] proposed a graph-matching feature selection method that preserves the relationship between the predicted vectors and the target vectors and takes high-order graph-matching. Furthermore, Zhu et al. [13] proposed a new loss function based on matrix-similarity that not only considers the natural relationship of clinical scores and label, but also the spatial relationship of samples to take a better feature selection and classification. This method was also developed in [14]. In order to use multi-modal data more effectively, Shi et al. [15] proposed a method that fuses the features from different modalities by using the

* Corresponding author.
E-mail address: dlyyang@gmail.com (Y. Yang).

http://dx.doi.org/10.1016/j.neucom.2015.08.111
0925-2312/

pairwise coupled-diversity correlation. However, those previous methods based on $L_{2,1}-$norm regularizer suffer from some of three main disadvantages: (1) $L_{2,1}-$norm based loss function does not take into account different distances between target labels and prediction values; (2) $L_{2,1}-$norm based feature selection does not possess sufficient flexibility to adapt to different types of data sources and select more informative features; (3) intrinsic correlation between the processes of feature selection and classification (or regression) is inevitably ignored in existing methods.

In this paper, we propose a novel loss function that combines the $L_{2,p}-$norm of prediction loss function and $L_{2,q}-$norm of feature selection, and utilizes a joint model to conduct the feature selection and classification simultaneously. We also introduce a new item that keeps local structure information between samples in the feature space and prediction values in the projected space. The $L_{2,p}-$norm of prediction loss function attempts to adjust distances between predict values and target labels, and controls distances at the point of convergence of loss function. The larger p is, the less widely the distances vary. The $L_{2,q}-$norm of feature selection tries to control the sparsity of feature selection. The larger q is, the less sparse the feature selection is. We can flexibly select appropriate p, q according to the data and thus achieve a better classification. The new spatial information item preserves relationships between samples, *i.e.* if two samples are close in original feature space, they are still close neighbors in the projected space. Experiments on the ADNI dataset have showed that our proposal indeed helps us to enhance the performance of disease status classification, comparing the state-of-the-art methods.

2. Materials and preprocessing

In this paper, we use the ADNI (Alzheimer's Disease Neuroimaging Initiative) dataset[1] to evaluate our method. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). These data contains 202 samples including 51 samples for AD, 52 samples for NC, and 99 samples for MCI. We downloaded the MRI, PET and CSF data from the public ADNI website. Then we extract 93 features from MRI and 93 features from PET as well as 3 features from CSF following the widely used procedures such as in [3,16]. The detailed statistic information of these samples is showed in Table 1. Note that the numbers in this table represent the subjects' number or values' range corresponding to each category. MCI-C represents MCI Converters and MCI-NC represents MCI Non-Converters.

Table 1
Statistic information of samples in our dataset.

Items	AD (51)	NC (52)	MCI-C (43)	MCI-NC (56)
Female/male	18/33	18/34	15/28	17/39
Age	75.2 ± 7.4	75.3 ± 5.2	75.8 ± 6.8	74.8 ± 7.1
Education	14.7 ± 3.6	15.8 ± 3.2	16.1 ± 2.6	15.8 ± 3.2
MMSE	23.8 ± 2.0	29.0 ± 1.2	26.6 ± 1.7	27.5 ± 1.5
ADAS-Cog	18.3 ± 6.0	7.4 ± 3.2	12.9 ± 3.9	10.2 ± 4.3

3. Our method

In this section, we describe our feature selection and classification framework for AD/MCI diagnosis. Given the MRI, PET, and CSF features, we construct the feature matrix with each column concatenates these multi-modal features, and the target matrix or vector of ground truth with each column representing a sample that concatenates a class label and two clinical scores (ADAS-Cog and MMSE) or contains only a class label.

3.1. Preliminaries

Let $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]\in\mathbf{R}^{\mathbf{d}\times\mathbf{n}}$ and $\mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_n]^T\in\mathbf{R}^{\mathbf{c}\times\mathbf{n}}$, where n, d and c denote the number of samples, dimension of features, and number of target values, respectively. In our work, the target values correspond to a class label and two clinical scores, thus c can be equal to 1, 2 or 3. Commonly, we can predict the target variables by preforming a linear transformation for features, which is formulated as the following equation:

$$f(\mathbf{X})=\mathbf{X}^T\mathbf{W}=\hat{\mathbf{Y}} \tag{1}$$

where $\mathbf{W}\in\mathbf{R}^{d\times c}$ is a regression matrix and $\hat{\mathbf{Y}}$ is the predict target matrix. Each column in $\hat{\mathbf{Y}}$ corresponds to one of target variable and each row in $\hat{\mathbf{Y}}$ corresponds to one of samples. Note that, in this paper, $\mathbf{X}$ has been appended one additional dimension with value of 1 for every sample to include the bias item. Like the proposal of other literatures, if we restrict to select the same features to predict all target variables, we can formulate the feature selection method as follows:

$$\min_{\mathbf{W}} f(\mathbf{W})+\lambda\|\mathbf{W}\|_{2,1} \tag{2}$$

where $f(\mathbf{W})$ is the loss function between prediction values and target values depending on $\mathbf{W}$ and $\|\mathbf{W}\|_{2,1}=\sum_{i=1}^{d}\|\mathbf{W}^i\|_2$, $\mathbf{W}^i$ is the ith row of $\mathbf{W}$, λ is a weight parameter. The $L_{2,1}-$norm regularizer $\|\mathbf{W}\|_{2,1}$ let the model simultaneously select or not select a feature for predicting all target variables. Specifically, the L_2-norm regularizer in each row of $\mathbf{W}$ enforces all tasks to select the same features, and the L_1-norm regularizer imposes the constraint of sparseness in the feature selecting stage to select the most important and discriminative features. In our classification problem, our $L_{2,q}-$norm is like this $L_{2,1}-$norm but with additional flexibility and adaptability.

The loss function $f(\mathbf{W})$ in Eq. (2) is commonly defined as the distance between target values $\mathbf{Y}$ and predicted values of all samples, which is presented as follows:

$$\begin{aligned} f(\mathbf{W}) &= \|\mathbf{Y}-\mathbf{X}^T\mathbf{W}\|_F^2 \\ &= \|\mathbf{Y}-\hat{\mathbf{Y}}\|_F^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{c}(\mathbf{y}_{ij}-\mathbf{y}_{ij})^2 \end{aligned} \tag{3}$$

This distance based similarity metric has been proved effective and efficient in many literatures such as [3,13,17]. The lower this function value is, the more accurate the prediction is.

3.2. The proposed method

As described in the introduction, most $L_{2,1}-$norm based methods have three main disadvantages. One is that the $L_{2,1}-$norm of prediction loss item does not take into account different distances between target labels and prediction values. Another is that the $L_{2,1}-$norm of feature selection is not flexible enough to select more useful features. The third is that they usually first

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

conducted feature selection and then use these selected features to train another classifier (*i.e.* SVM) or regression model for the AD diagnosis. Although the used classifier or regression model is possibly more effective, the intrinsic correlation between the processes of feature selection and classification or regression is inevitably ignored. The selected features using Eq. (2) may be not suitable for other regression or classification models. To overcome these disadvantages, we propose a method which incorporates additional flexibility and adaptability by employing the more generalized $L_{2,p}-$norm of loss function and $L_{2,q}-$norm of feature selection and utilizes a joint model to perform feature selection and classification simultaneously. We also introduce a new constraint that keeps local spatial information between samples in the feature space and prediction values in the projected space. Our newly devised loss function is shown as below:

$$\min_{\mathbf{W}} \|\mathbf{Y}-\mathbf{X}^T\mathbf{W}\|_{2,p}+\lambda\|\mathbf{W}\|_{2,q}+\beta\ \mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \tag{4}$$

where $\|\mathbf{Y}-\mathbf{X}^T\mathbf{W}\|_{2,p}=\sum_i\|(\mathbf{Y}-\mathbf{X}^T\mathbf{W})^i\|_2^p$, $\|\mathbf{W}\|_{2,q}=\sum_i\|\mathbf{W}^i\|_2^q$ (for a matrix $\mathbf{M}$, $\mathbf{M}^i$ represents the ith row of $\mathbf{M}$) and $\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$ is the trace of matrix $\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}$. λ and β are the weight of item $\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$ and item $\|\mathbf{W}\|_{2,q}$, respectively. $\mathbf{L}$ in the item $\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$ is the Laplacian matrix [18]. $\mathbf{L}$ can be constructed as the following short description.

Firstly, we construct a distance matrix $\mathbf{A}$ between samples in the feature space. For each sample in the dataset, we maintain m minimal distances values and set others to be 0. In our method, we set $m=5$. Then we construct a graph $\mathbf{G}$ that its vertexes are samples and weights of edge between two vertexes are Gaussian Kernel distance of corresponding samples. The zero value in $\mathbf{A}$ represents that there is no edge connecting the corresponding two samples. When the graph is constructed, we can compute the Laplacian matrix $\mathbf{L}$ of this graph according to its definition.

In our loss function, the $L_{2,p}-$norm item $\|\mathbf{Y}-\mathbf{X}^T\mathbf{W}\|_{2,p}$ attempts to adjust distances between predict values and target labels. If $p>1$, then the large distances will get larger and they are punished more heavily than the smaller ones are. Thus, this will push their prediction values closer to target values and when the loss function converges, the distances of all samples do not vary too widely. While $p<1$, the situation is on the contrary. The $L_{2,q}-$norm item $\|\mathbf{W}\|_{2,q}$ tries to select more informative features by controlling the sparsity of feature selection. The larger q is, the less sparse the feature selection is. We can flexibly select appropriate p, q according to the data and thus achieve a better classification.

The last item $\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$ considers spatial structure of samples, which makes relative distance of nearby samples maintained in the projected space. Note that

$$\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})=\frac{1}{2}\sum_{k=1}^{c}\sum_{i,j=1}^{n}a_{i,j}\left(\left(\mathbf{W}^T\mathbf{X}\right)_{k,i}-\left(\mathbf{W}^T\mathbf{X}\right)_{k,j}\right)^2 \tag{5}$$

where n is the number of samples, c is the dimension of target values and $a_{i,j}$ is the weight in the graph $\mathbf{G}$. Hence, we can guarantee that the m most nearby samples in feature space are also nearby in the projected space in the process of minimizing our $\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$. Therefore, we can reduce mistakes classifying closed samples into two classes and achieve a better classification.

3.3. Optimization

In this part, we devise an effective and efficient algorithm for solving the minimization problem in Eq. (4). Let $\mathbf{R}=\mathbf{Y}-\mathbf{X}^T\mathbf{W}=[\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_n]^T$, and $\mathbf{D}^r$, $\mathbf{D}^l$ are diagonal matrices with their diagonal elements $\mathbf{D}^l_{ii}=\frac{1}{\frac{2}{p}\|\mathbf{r}_i\|_2^{2-p}}$, $\mathbf{D}^r_{jj}=\frac{1}{\frac{2}{q}\|\mathbf{w}_j\|_2^{2-q}}$ respectively. Then Eq. (4) can be transformed to Eq. (6):

$$\min_{W}\mathrm{tr}\left((\mathbf{Y}-\mathbf{X}^T\mathbf{W})^T\mathbf{D}^l(\mathbf{Y}-\mathbf{X}^T\mathbf{W})\right)+\lambda\ \mathrm{tr}(\mathbf{W}^T\mathbf{D}^r\mathbf{W})+\beta\ \mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}) \tag{6}$$

The derivative of Eq. (6) with respect to $\mathbf{W}$ is difficult to compute, but we can consider that $\mathbf{D}^l$ and $\mathbf{D}^r$ are constant terms since we assume that we have already known $\mathbf{W}$ in the last iteration. Then we can compute the derivative of Eq. (6) with respect to $\mathbf{W}$. By setting the derivative to be zero, we can get the following equation

$$-\mathbf{X}\mathbf{D}^l(\mathbf{Y}-\mathbf{X}^T\mathbf{W})+\lambda\mathbf{D}^r\mathbf{W}+\beta\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}=0 \tag{7}$$

Then we can get new value of $\mathbf{W}$ computed as

$$\mathbf{W}=(\mathbf{X}\mathbf{D}^l\mathbf{X}^T+\lambda\mathbf{D}^r+\beta\mathbf{X}\mathbf{L}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{D}^l\mathbf{Y} \tag{8}$$

We start with a randomly initialized $\mathbf{W}$ and compute $\mathbf{D}^l$ and $\mathbf{D}^r$, then update $\mathbf{W}$ according to Eq. (8). We keep repeating this process until the value of the objective function in Eq. (6) converges. The overall of our algorithm is described as below:

Algorithm 1. Pseudo code for solving the problem in (4)

Algorithm 1: Pseudo code for solving the problem in (4)

Input : $\mathbf{Y}\in\mathbf{R}^{n\times c}$, $\mathbf{X}\in\mathbf{R}^{d\times n}$, p, q, λ, β;
Output: $\mathbf{W}$;
Initialize $t=0$, $\mathbf{W}(t)$ as a random matrix;
Construct Laplacian matrix L;
repeat
  Calculate $\mathbf{D}^l$ and $\mathbf{D}^r$;
  Update $\mathbf{W}(t+1)$ through Eq. (8);
  $t\leftarrow t+1$;
until *there is no change to* $\mathbf{W}(t)$;
return $\mathbf{W}$;

4. Experimental results and analysis

4.1. Experimental settings

We evaluate our proposed method by comparing the performance of solving three binary classification problems between AD and NC, MCI and NC, as well as MCI-C and MCI-NC on the ADNI dataset. In the classification between MCI and NC, MCI-C samples and MCI-NC samples were merged with the label of MCI. Even though the proposed method can simultaneously predict ADAS-Cog and MMSE scores as well as the class label, we only focus on the classification problem because the prediction of ADAS-Cog and MMSE scores are not our main task, they are just the sub-tasks that help us do a better classification.

We use the multi-task formulation by concatenating both the class label and clinical scores (ADAS-Cog or MMSE) to form the target matrix. For very classification problem, we utilize the single-modality features such as MRI or PET, and the multi-modality features such as MP (MRI and PET) or MPC (MRI, PET and CSF) respectively to train our feature selection and classification model. After training, we leverage Eq. (1) to predict the class label.

We use four metrics of ACC (accuracy), SEN (sensitivity), SPE (specificity), and AUC (area under curve) to evaluate the classification performance of all compared methods. In order to get more reliable results, We employ the 10-fold cross-validation to evaluate all compared methods.

4.2. Parameters selecting

In our proposed method, there are totally four parameters, namely p, q, λ, β. We respectively set $p \in \{0.2, 0.4, \ldots 1.8\}, q \in \{0, 0.2, 0.4, \ldots 1.8\}$ and $\lambda \in \{10^{-1}, 1, \ldots 10^{5}\}$, $\beta \in \{10^{-4}, 10^{-3}, \ldots 10^{2}\}$ in our experiments. Fig. 1 shows respective accuracy results when choosing different parameters in the classification of AD *vs.* NC, MCI *vs.* NC, and MCI-C *vs.* MCI-NC. Different accuracy values are labeled by different colors. Each block in left pictures represents the best accuracy value when setting a pair of fixed parameters p and q while parameters λ and β are variable, and each block in right pictures represents the best accuracy value when setting a pair of fixed parameters λ and β while parameters p and q are variable. Vertical direction in left pictures represents parameter p and horizontal direction represents q. Similarly, vertical direction in right pictures represents parameter λ and horizontal direction represents β.

From Fig. 1, we can see that in all three classification problems, p is not equal to 2 and q is not equal to 1 when we get the best accuracy results. Thanks to the more generalized $L_{2,p}$-norm based prediction loss function and $L_{2,q}$-norm based feature selection, we can choose the best p and q according to different data and different task flexibly and adaptively. Specifically, in the classifications of AD

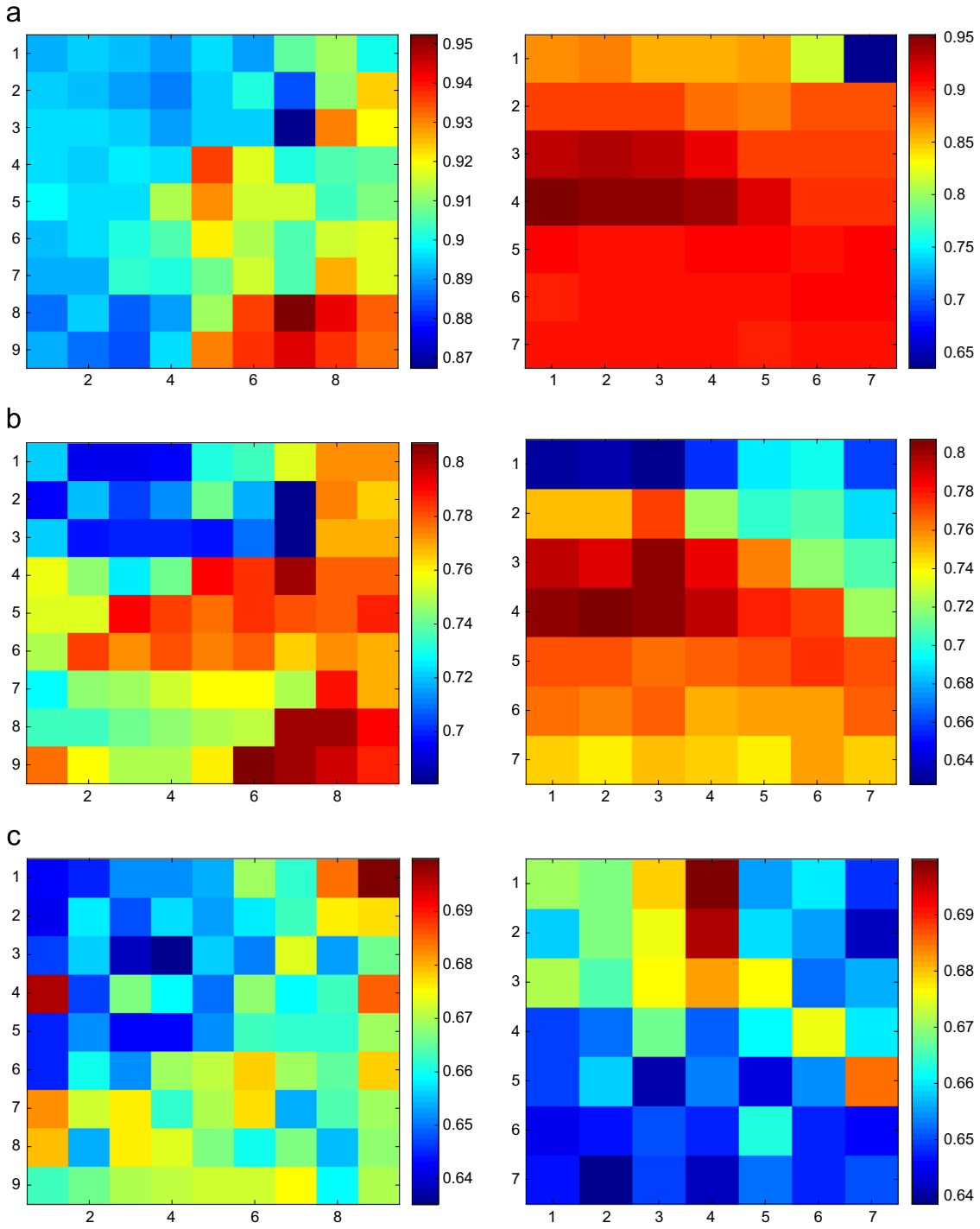


Fig. 1. Accuracy results of classifications when choosing different parameters. (For interpretation of the references to color in the text, the reader is referred to the web version of this paper.)

vs. NC and MCI *vs.* NC, we get better accuracy results when p and q locate in the bottom right area. More precisely, Fig. 1 shows when p is on the range of 1.6 and 1.8 and q is on the range of 1.2 and 1.6, our method can achieve the best performance. In the classification of MCI-C *vs.* MCI-NC, we can get the best result when p and q locate in the up right corner with p is equal to 0.2 and q is equal to 1.8. In the classification of MCI-C *vs.* MCI-NC, we get the best result when p is much smaller camparing to the classification of AD *vs.* NC or MCI *vs.* NC. This may because the MCI-C *vs.* MCI-NC is difficult to classified, so the distances between predict values and target values vary more widely. The larger p in the item $L_{2,p}-$norm item $\left\|\mathbf{Y}-\mathbf{X}^T\mathbf{W}\right\|_{2,p}$ makes the distances between predict values and target values vary not too widely when the loss function converges. The larger q in the item $\lambda\|\mathbf{W}|_{2,q}$ makes feature selection less sparse and on the contrary, the smaller q makes feature selection sparser and the number of selected features is smaller. We can see that optimal value of q in the classification of AD *vs.* NC and MCI *vs.* NC is smaller than that in MCI-C *vs.* MCI-NC. This is because the symptoms of people with AD or MCI are obvious different from NC ones, while the difference of symptoms between MCI-C and MCI-NC is less distinct. So AD *vs.* NC and MCI *vs.* NC are more easily classified and some features contain much discriminated information. This makes the feature selection sparser and leads to smaller q. In the classification of MCI-C *vs.* MCI-NC, we get the best result when p is much smaller camparing to the classification of AD *vs.* NC or MCI *vs.* NC. This may because the MCI-C *vs.* MCI-NC is difficult to classified, so the distances between predict values and target values vary more widely. We need to select appropriate p and q according to the data and classification task.

Let us take a look at λ and β. In the classifications of AD *vs.* NC and MCI *vs.* NC, when λ is 10 or 100 our results are best. This means the process of feature selection is very important for better classification. In the classification of MCI-C *vs.* MCI-NC, the value of λ is a little small. As the aforementioned reason that the difference of symptoms between MCI-C and MCI-NC is less distinct and have not very discriminated features, the feature selection is less import when comparing the other two classification problems. When our results are best, β in item $\beta\,\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$ is on the range of 0.0001 and 0.1. This means the constraint of projective value between samples is not as important as better feature selection in our data. But we also can see that $\beta\,\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$ is also indispensable in our classification tasks. We can easily see that in the classifications of MCI *vs.* NC and MCI-C *vs.* MCI-NC, β at the left of its optimal location makes the accuracy of classification lower when λ is invariable. Although the item $\beta\,\mathrm{tr}(\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W})$ is not important than the item $\lambda\|\mathbf{W}\|_{2,q}$, it is also effective for improving classification accuracy.

4.3. Competing methods

We choose other eight methods to compare with our proposed method, which are listed below:

- LDA+kNN and RF (random forest): We use these basic feature selection and classification methods as the baseline in order to demonstrating how hard the task is. In the LDA+kNN method, we first use LDA (Linear Discriminant Analysis) to map the original feature to one dimension, then use the kNN to do classification. The number of neighborhood k is 5. In the random forest classifier, we use 50 decision trees.
- Original features based method: We choose this method that uses original features directly rather than the selected ones to conduct the classification to show the importance of feature selection in our task. This method was denoted with the suffix "N".
- High-Order Graph Matching method (HOGM) [4]: It is a Single-task method whose main contribution refers to using a high-Order relationship of samples between the predicted vectors and target vectors. This method does the classification using SVM after the feature selection, but ours does simultaneously the feature selection and classification. We choose this method to consider the performance of our method using a class label and two clinical scores.
- Matrix-Similarity Based method (MS-S) [14]: This method uses a new loss function based on matrix-similarity that not only considers the natural relationship of clinical scores and label, but also the spatial relationship of samples. Because of considering these sophisticated relationships, we can take a better feature selection, and as a result, improve the classification performances in AD/MCI diagnosis. The suffix "S" denotes performing the single task of classification.
- Joint Coupled-Feature Representation and Coupled Boosting method (JCFCB) [15]: We use this method to compare the performance of its Multi-Modal approach with ours. Firstly, this method computes the feature representation using intra-coupled and inter-coupled interaction relationship. Secondly, the features from different modalities are fused for classification by leveraging the pairwise coupled-diversity correlation.
- M3T [3]: This is a Multi-Modal Multi-Task method and we use it to compare with our proposed method that utilizes Multi-Modal Multi-Task framework. This method firstly selects the subset of features using Multi-Task method from each modality, then using the multi-modal support vector for the classification of AD and MCI.
- Manifold regularized Multi-Task Feature Selection (M2TFS) [19]: This is another Multi-Modal Multi-Task method used to compare. This method uses the $L_{2,1}-$norm regularizer and a manifold based regularizer for Feature Selection. By the manifold based regularizer, geometric information in each modality can be preserved. Each task corresponds to the classification on each modality. In our experiments, M2TFS-C represents concatenating features from multi-modality for classification and M2TFS-K represents using multiple kernels to fuse multi-modality features.

4.4. Classification results

Table 2 is the classification results for all compared methods. From these results, we can see that our proposed method enhances the classification performance when compared the other state-of-art methods in many classification problems. In addition to this, more details of our experimental results can be found:

1. The process of feature selection is very important for classification. It can be seen clearly that the baseline method without feature selection process has the worst performance in all three classification problems, while others with feature selection process get significantly better results.
2. Using Multi-modality information can actually improve the performance of classification. As the results shown in Table 2 and Fig. 2, all the results of methods using Multi-modality features are better than the same ones using single-modality feature. For example, the methods using MP get better results than those using MRI or PET. Meanwhile, using MPC is also better than using MP.
3. Although using multi-task can usually improve classification performance, but results are not always so. Whether using multi-task can improve classification performance or not is depended on the relationship between class label and sub-task values. From Table 1 we can find that in the classifications of AD *vs.* NC and MCI *vs.* NC, ADAS-Cog and MMSE scores are very discriminative for those three classes, while in the classification of MCI-N *vs.* MCI-NC, these two scores are not very discriminative. So in the two classification tasks in front, as shown in Fig. 2, adding two sub-tasks of predicting ADAS-Cog and MMSE scores can help enhance the classification

Table 2
The ACC, SEN, SPE, and AUC (%) results of all the compared methods.

Feature	Method	AD vs. NC				MCI vs. NC				MCI-C vs. MCI-NC			
		ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC	ACC	SEN	SPE	AUC
MRI	LDA+*k*NN	59.5	57.8	59.9	56.2	60.6	67.6	48.4	58.2	53.3	49.6	54.7	49.5
	RF	86.8	86.6	87.2	87.0	69.5	85.6	41.5	65.2	53.3	35.5	69.8	51.2
	MRI-N	89.5	82.7	86.3	95.3	68.3	92.6	39.2	82.5	60.2	15.5	92.3	68.7
	HOGM	**93.4**	89.5	92.5	97.1	77.7	**95.6**	51.4	84.4	66.8	36.7	95.0	72.2
	MS-S	91.2	85.9	92.5	96.7	76.7	93.3	37.6	83.7	64.5	24.9	**95.8**	70.6
	M3T	92.6	87.2	95.9	97.5	78.1	94.5	54.0	83.1	67.1	37.7	92.0	72.5
	Proposed	92.9	**86.9**	**96.5**	**96.2**	**78.3**	94.5	**54.8**	**82.4**	**66.7**	**45.4**	86.9	**71.9**
PET	LDA+*k*NN	51.8	50.1	49.9	52.5	53.4	57.1	44.0	48.6	48.0	51.3	44.6	47.2
	RF	81.1	82.0	81.7	81.6	68.4	90.6	29.0	63.9	54.1	41.5	65.6	52.5
	PET-N	86.2	83.5	84.8	94.8	69.0	95.0	30.8	77.9	62.2	21.6	93.1	71.3
	HOGM	91.7	91.1	92.8	95.6	74.7	**96.5**	43.2	79.3	66.6	35.5	**95.5**	72.4
	MS-S	87.9	85.7	90.9	94.7	73.8	**96.5**	36.2	78.7	65.1	31.0	**95.5**	73.5
	M3T	90.9	90.5	93.1	96.4	77.2	94.5	44.3	80.5	67.0	39.1	93.2	73.1
	Proposed	**90.8**	**84.1**	**93.9**	**93.6**	**76.8**	94.3	**50.3**	**80.7**	**67.4**	**58.6**	78.6	**73.4**
MP	LDA+*k*NN	78.9	79.6	78.1	77.6	54.7	57.8	48.9	54.3	57.9	52.7	62.5	57.6
	RF	86.8	89.5	82.8	86.5	71.0	88.8	39.7	62.5	53.9	37.7	67.8	54.2
	MP-N	89.7	92.2	85.9	96.1	71.6	96.1	43.9	82.7	62.7	22.6	93.5	73.2
	M2TFS-C	91.0	90.4	91.4	95.0	73.4	76.5	67.1	78.0	58.4	52.3	63	60.0
	M2TFS-K	95.0	**94.9**	95	97.0	79.3	85.9	66.6	82.0	68.9	64.7	71.8	70.0
	HOGM	**95.2**	92.8	95.4	97.8	79.5	**96.6**	58.6	84.6	67.6	45.5	**96.8**	75.1
	MS-S	90.8	92.6	93.8	96.7	76.3	97.0	39.9	83.4	66.9	33.9	96.0	75.7
	M3T	94.0	92.0	96.3	98.0	78.4	95.0	57.7	83.9	67.9	47.0	93.3	75.7
	JCFCB	94.7	94.2	96.9	93.1	80.1	81.7	**76.2**	75.7	–	–	–	–
	Proposed	94.4	94.2	**95.7**	**97.6**	**79.3**	95.7	56.1	**83.6**	**69.2**	**47.3**	96.2	**76.8**
MPC	LDA+*k*NN	82.5	84.2	81.5	81.6	55.4	58.7	50.1	55.4	58.8	55.3	61.7	58.7
	RF	89.4	90.9	88.2	91.8	73.7	90.1	45.4	67.6	55.6	41.3	69.6	56.5
	MPC-N	90.8	93.1	88.3	96.5	72.5	96.3	47.1	84.1	64.1	23.1	93.6	73.9
	HOGM	**95.6**	94.5	96.9	98.5	80.6	96.7	64.7	86.2	68.8	47.5	**98.5**	75.3
	MS-S	92.5	94.1	93.8	97.6	77.1	**97.1**	47.5	83.9	67.8	34.1	96.8	75.8
	M3T	94.6	93.1	96.4	98.5	80.1	95.2	58.7	84.3	68.5	47.5	92.7	76.0
	Proposed	**95.3**	**94.6**	**96.5**	**98.7**	**80.7**	96.3	**68.2**	**87.2**	**69.8**	**48.7**	96.7	**77.5**

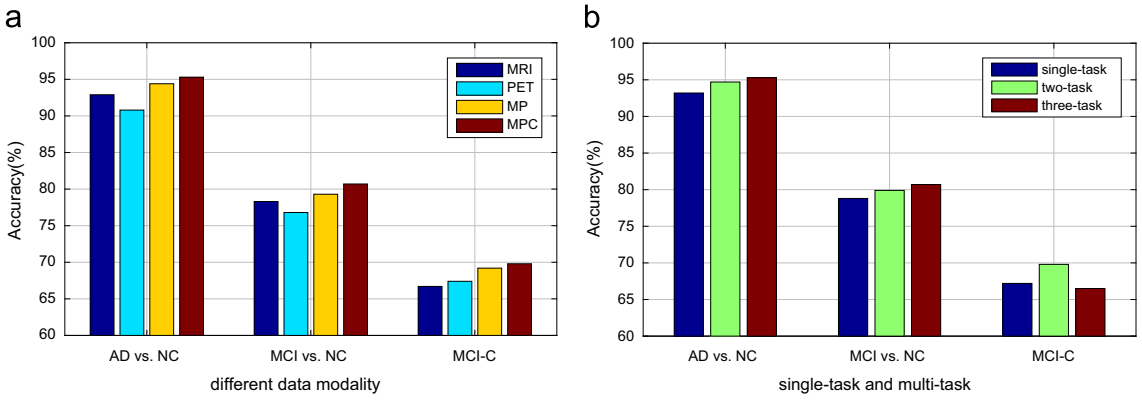


Fig. 2. Accuracy comparison. (a) Accuracy comparison between different data modality in different classification task and (b) accuracy comparison between single-task and multi-task in different classification task using MPC data modality.

performance. But in the last classification task, adding those two scores without distinction can get even worse result.

5. Conclusion

In this paper, we considered three binary classification problems of AD *vs.* NC, MCI *vs.* NC, as well as MCI-C *vs.* MCI-NC in AD diagnosis. Because most recent studies of $L_{2,1}$−norm method on AD diagnosis have showed three main limitations, we propose a method that does the $L_{2,p}$−norm of prediction loss function and $L_{2,q}$−norm of feature selection, as well as does the feature selection and classification simultaneously. We also introduce a new constraint that keeps local structure information between samples in the feature space and prediction values in the projective space. The $L_{2,p}$−norm of prediction loss function attempts to adjust distances between predict and target values, and controls distances at the point of convergence of loss function. The larger p is, the less widely distances vary. The $L_{2,q}$−norm of feature selection tries to control the sparsity of feature selection. The larger q is, the less the sparse feature selection is. We can select appropriate p, q according to the data and to do a better classification. The new relational information constraint keeps relationships between samples that makes closed samples in feature space also closely projected. At the last we conducted many experiments on the ADNI dataset, and it showed that our newly proposed method enhanced the performances of disease status classification, compared to the state-of-the-art methods.

Acknowledgements

This work was supported in part by the National Nature Science Foundation of China under Project 61572108.

Data collection and sharing for ADNI dataset was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, H.M. Arrighi, Forecasting the global burden of Alzheimer's disease, Alzheimer's Dement. 3 (3) (2007) 186–191.

[2] R.C. Petersen, R. Doody, A. Kurz, R.C. Mohs, J.C. Morris, P.V. Rabins, K. Ritchie, M. Rossor, L. Thal, B. Winblad, Current concepts in mild cognitive impairment, Arch. Neurol. 58 (12) (2001) 1985–1992.

[3] D. Zhang, D. Shen, A.D.N. Initiative, et al., Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease, Neuroimage 59 (2) (2012) 895–907.

[4] F. Liu, H.-I. Suk, C.-Y. Wee, H. Chen, D. Shen, High-order graph matching based feature selection for alzheimer's disease identification, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer, 2013, pp. 311–318.

[5] H. Wang, F. Nie, H. Huang, S. Risacher, A.J. Saykin, L. Shen, et al., Identifying adsensitive and cognition-relevant imaging biomarkers via joint classification and regression, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011, Springer, 2011, pp. 115–123.

[6] X. Zhu, Z. Huang, H.T. Shen, J. Cheng, C. Xu, Dimensionality reduction by mixed kernel canonical correlation analysis, Pattern Recognit. 45 (8) (2012) 3003–3016.

[7] M. Lustig, D. Donoho, J.M. Pauly, Sparse mri: the application of compressed sensing for rapid mr imaging, Magn. Reson. Med. 58 (6) (2007) 1182–1195.

[8] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.

[9] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, T.-S. Chua, Exploiting web images for semantic video indexing via robust sample-specific loss, IEEE Trans. Multimed. 16 (6) (2014) 1677–1689.

[10] Y. Yang, Y. Yang, Z. Huang, H.T. Shen, F. Nie, Tag localization with spatial correlations and joint group sparsity, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 881–888.

[11] X. Zhu, Z. Huang, Y. Yang, H.T. Shen, C. Xu, J. Luo, Self-taught dimensionality reduction on the high-dimensional small-sized data, Pattern Recognit. 46 (1) (2013) 215–229.

[12] H. Wang, F. Nie, H. Huang, S. Risacher, A.J. Saykin, L. Shen, et al., Identifying adsensitive and cognition-relevant imaging biomarkers via joint classification and regression, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011, Springer, 2011, pp. 115–123.

[13] X. Zhu, H.-I. Suk, D. Shen, A novel matrix-similarity based loss function for joint regression and classification in ad diagnosis, NeuroImage 100 (2014) 91–105.

[14] X. Zhu, H.-I. Suk, D. Shen, Matrix-similarity based loss function and feature selection for alzheimer's disease diagnosis, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 3089–3096.

[15] Y. Shi, H.-I. Suk, Y. Gao, D. Shen, Joint coupled-feature representation and coupled boosting for ad diagnosis, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014, pp. 2721–2728.

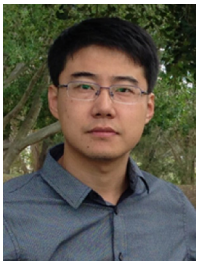
[16] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, A.D.N. Initiative, et al., Multimodal classification of alzheimer's disease and mild cognitive impairment, NeuroImage 55 (3) (2011) 856–867.

[17] H.-I. Suk, C.-Y. Wee, D. Shen, Discriminative group sparse representation for mild cognitive impairment classification, in: Machine Learning in Medical Imaging, Springer, 2013, pp. 131–138.

[18] R. Merris, Laplacian matrices of graphs: a survey, Linear Algebra Appl. 197 (1994) 143–176.

[19] B. Jie, D. Zhang, B. Cheng, D. Shen, Manifold regularized multi-task feature selection for multi-modality classification in alzheimer's disease, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Springer, 2013, pp. 275–283.

Mingxing Zhang received the B.S. degree in Computer Science from the Inner Mongolia University of Technology, Hohhot, China, in 2012. He is currently a postgraduate student in the School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests are image processing, computer version, and machine learning.



Yang Yang received the B.S. degree from Jilin University, Changchun, China, in 2006, the M.E. degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from the University of Queensland, Brisbane, QLD, Australia, in 2013. Currently, he is a Professor at the School of Computer Science & Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include multimedia information retrieval, social media analysis, and machine learning.

Fumin Shen received his Bachelor degree at 2007 and Ph.D. degree at 2014 from Shandong University and Nanjing University of Science and Technology, China, respectively. Now he is a Lecturer at the University of Electronic Science and Technology of China. His major research interests include computer vision and machine learning, including face recognition, image analysis, hashing methods, and robust statistics with its applications in computer vision.

Hanwang Zhang received the B.Eng. (Hons.) degree in Computer Science from Zhejiang University, Hangzhou, China, in 2009, and the Ph.D. degree in Computer Science from the National University of Singapore, Singapore, in 2014. His main research interests are multimedia and computer vision, focusing on developing techniques for efficient search and recognition in image contents.

Dongxiang Zhang is a Senior Research Fellow in the School of Computing, National University of Singapore. He received the B.Sc. degree from Fudan University, China in 2006 and the Ph.D. degree from National University of Singapore in 2012. He worked as a Research Fellow at the NeXT research center in Singapore from 2012 to 2014. His research interests include spatial information retrieval, moving object queries and multimedia indexing.